

# TAB: Transformer Attention Bottlenecks enable User Intervention and Debugging in Vision-Language Models

## Supplementary Material

### A. Implementation details

#### A.1. Framework

We adopt our CC framework from CLIP4IDC [27] (see Fig. A1). We resize the input images to  $224 \times 224$ . The images are then patchified into 49 patches for B/32 and 196 patches for B/16 using the first convolution layer in CLIP ViT-B/32 and ViT-B/16, respectively. The intermediate embedding dimension in the vision Transformers is  $d = 768$  and the final projected embedding is 512. We use the encoder-decoder language model to perform the next token prediction.

#### A.2. Training

We fix the first convolution layer and use the retrieval loss (Eq. (10)) to align vision and language blocks. We use the initial learning rate of  $10^{-7}$  and use a cosine scheduler with Adam optimizer. We continue the training in the alignment stage for 12 epochs (see Tab. A1). We drop the retrieval loss and the text encoder from the framework for the text generation stage and connect the pretrained vision Transformer to the language model. We use Cross Entropy loss and leverage the groundtruth box annotations to supervise the activation map in the bottleneck. We train the text generation stage for 50 epochs with Bert’s implementation of Adam optimizer settings (see Tab. A2).

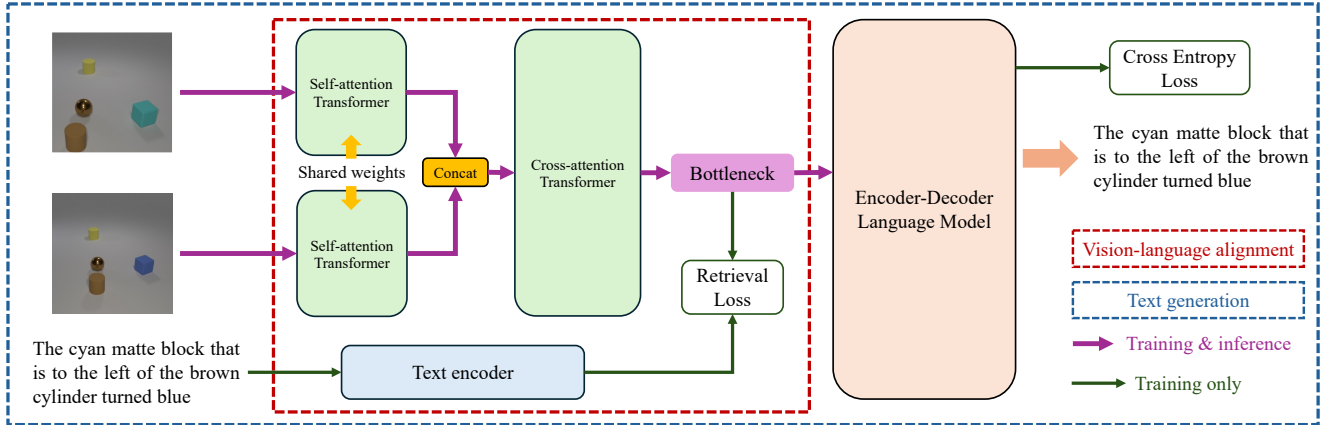


Figure A1. We use a two-stage training method for Image Difference Captioning. First, the visual embeddings extracted from the image pair in the bottleneck are aligned with the textual embeddings of the captions. In the second stage, we use a Cross Entropy loss to predict the next token.

Table A1. Vision-language alignment recipe for TAB4IDC.

Config	Value
Optimizer	Adam
Base LR	$1e^{-4}$
Scheduler	cosine decay [45]
Weight decay	0.2
Momentum	$\beta_1 = 0.9, \beta_2 = 0.98$
epsilon	$1e^{-6}$
Batch size	128
Warmup proportion	0.1
Training epochs	12

Table A2. Text generation recipe for TAB4IDC.

Config	Value
Optimizer	Adam
Base LR	$1e^{-4}$
Scheduler	linear decay
Weight decay	0.01
Momentum	$\beta_1 = 0.9, \beta_2 = 0.999$
epsilon	$1e^{-6}$
Batch size	64
Warmup proportion	0.1
Training epochs	50
Max words	32

## B. Attention visualization details

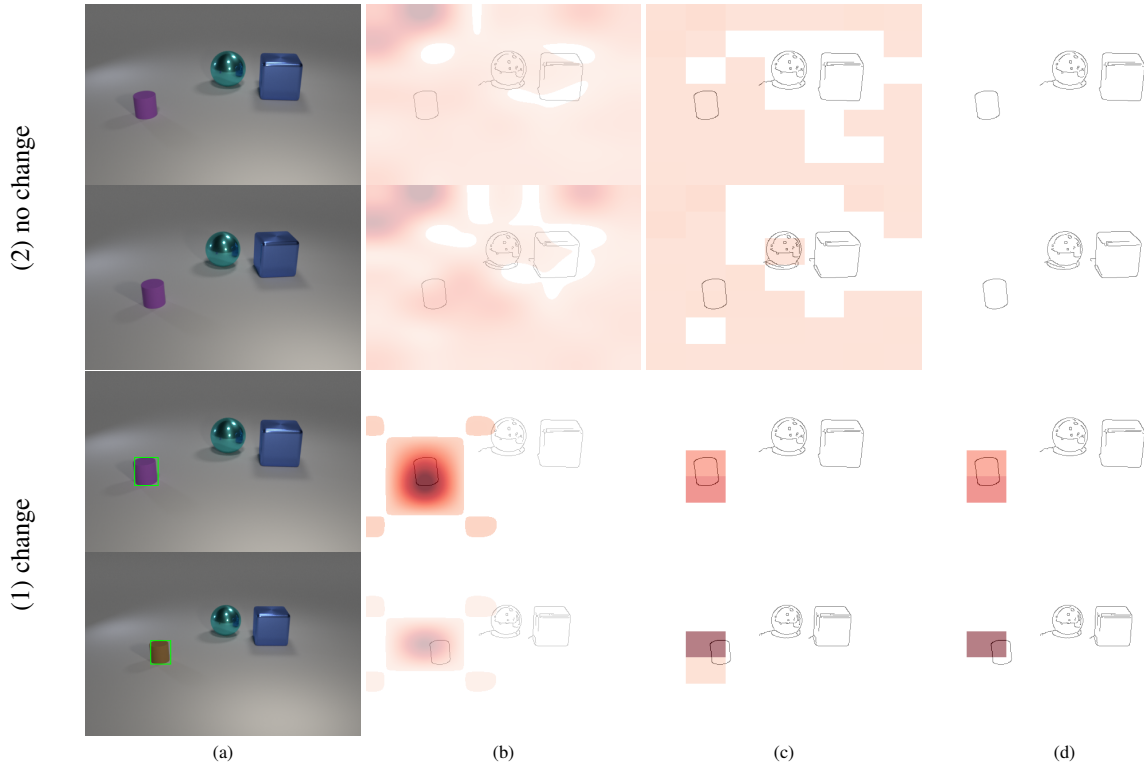


Figure A2. For each input pair in (a), the interpolation method in CLIP4IDC [27] (see Fig. A3) yields many nonzero attention values for no-change pairs (2b). In our visualization, we use the nearest-neighbor method (Fig. A4) that leads to fewer nonzero values on no-change (2c). For baseline methods, the thresholding substantially improves the attention map’s quality for no-change pairs (2d), where there is no target object for detection.

Table A3. The common interpolation method used for upscaling the attention map leads to low  $PG^+$  accuracy (0.0%) on (a), here for CLIP4IDC [27]. We use the nearest neighbor approach (Fig. A4) and then threshold the attention map at different values to find the most important patches based on their values. This improves the baseline method’s overall  $PG^+$  (90.15%).

Method	ViT	Interpolation	Thresh.	Change	No-change	Mean
Baseline	B/32	cubic	✗	79.98	0.0	39.99
Ours	B/32	nearest neighbor	✓	84.10	96.21	<b>90.15</b>

```
import cv2
import numpy as np

def resize_map(attention_map=None, input_size=(480, 320)):
    resized_map = cv2.resize(attention_map.astype(np.uint8), input_size, interpolation=cv2.INTER_CUBIC)

    return resized_map
```

Figure A3. Prior works [27, 55] use cubic interpolation to resize the attention maps that lead to smooth edges in the heatmap. Yet, it also results in peak values over image patches that have near zero attention values.

```
import cv2
import numpy as np

def resize_map(attention_map=None, input_size=(480, 320)):

    resized_map = cv2.resize(attention_map.astype(np.float32), input_size, interpolation=cv2.
        INTER_NEAREST)

    return resized_map
```

Figure A4. In our visualization paradigm, we replace the interpolation method with the nearest neighbor such that the resized attention has fewer nonzero values.

### C. Generating groundtruth captions for OpenImages-I

We have access to all the inpainted object names in OpenImages-I [51], and we generate multiple sentences describing a particular change using templates listed in Tab. A4.







Pair type	Caption template
Add	the ... has appeared
	the ... has been newly placed
	the ... has been added
Drop	the ... has disappeared
	the ... is missing
	the ... is gone
	the ... is no longer there
No-change	no change was made
	there is no change
	the two scenes seem identical
	the scene is the same as before
	the scene remains the same
	nothing has changed
	nothing was modified
	no change has occurred
	there is no difference

Table A4. We follow [55] and use the caption templates to generate groundtruth change captions.

## D. Editing the attention values in MHSA layer does not yield a different caption



To further investigate the role of our proposed 1-head attention in the [intervention](#) task, we also supervise the MHSA attention in CLIP4IDC during the training and evaluate it similarly to Sec. 5.3. Compared to TAB, the MHSA layer does not yield a cause-and-effect relation with the predictions Tab. A5, perhaps due to information leakage in the architecture.





Table A5. Supervising 12 heads in the MHSA layer of CLIP4IDC does not enable (✗) user intervention as in TAB (✓)

Dataset	Attention 	Acc. Change		Acc. No-change		Acc. object name	
		base		base		base	
MHSA 	ZERO	99.97	99.97 ✗	100.0	100.0	89.78	89.76 ✗
	CORRECT ↑	99.97	99.97 ✗	100.0	100.0	89.78	89.80 ✓
TAB	ZERO	99.93	0.0 ✓	100.0	100.0 ✓	88.92	0.0 ✓
	CORRECT ↑	99.93	100.0 ✓	100.0	100.0 ✓	88.92	91.49 ✓
MHSA 	ZERO	95.00	95.0 ✗	99.17	99.17 ✗	-	-
	CORRECT ↑	95.00	95.0 ✗	99.17	99.17 ✗	-	-
TAB	ZERO	94.30	0.0 ✓	99.42	100.0 ✓	-	-
	CORRECT ↑	94.30	100.0 ✓	99.42	100.0 ✓	-	-

## E. Zeroshot change **localization**

### E.1. TAB is a better zero-shot change localizer compared to the MHSA layers in VLM captioners

We use the VLMs trained for CC on , which contain only one change, and aim to evaluate if the MHSA layer in CLIP4IDC [27] and TAB in TAB4IDC can localize multiple changes in an unseen dataset () without further training.

**Experiment** We include a SotA change detection (CD) method, CYWS [62], as an upper bound, and report the  $PG^+$  on  as a baseline accuracy for CYWS, TAB4IDC and CLIP4IDC [27]. CYWS [62] is a U-Net-based CD framework trained on COCO-Inpainted , using the bounding box supervision. We consider CYWS [62] an upper bound accuracy because it is trained on real-world image pairs with multiple changes () similar to .

















**Results** Overall, TAB outperforms the MHSA layer of CLIP4IDC [27] by  $\sim +44$  points in mean  $PG^+$  (Tab. A6). The main reason is TAB’s better performance on no-change pairs. On average, TAB and the MHSA layer are worse than CYWS [62] because they use the attention values (softmax output) to localize multiple changes in . We hypothesize that the large gap is due to attention values spreading over many changes (see Fig. A5), which causes them not to pass the heatmap discretization in  $PG^+$ . On , TAB consistently outperforms CYWS [62] (99.19 vs. 89.76%). Compared to the MHSA layer in CLIP4IDC [27], TAB improves the baseline accuracy of a CD method on .

Table A6. **Zero-shot **localization****: TAB performs better under  $PG^+$  in zero-shot change **localization** in  than the MHSA layer in CLIP4IDC [27], which is trained on . TAB also has smaller delta with CYWS [62], the upper bound zero-shot **localization** accuracy on , than CLIP4IDC [27].

Method	Train	Thresh.	Change		No-change		Mean	
								
CYWS [62]		✗	100.0	99.91	0.0	0.0	50.0	49.95
CYWS [62]		✓	99.92	81.73	100.0	99.72	<b>99.96</b>	89.76
CLIP4IDC [27]		✓	74.9	24.55	12.6	83.74	43.75	54.14
TAB		✓	75.9	98.40	100.0	99.98	87.95 (+44.2)	<b>99.19</b>

## F. STD additional results

### F.1. Localization

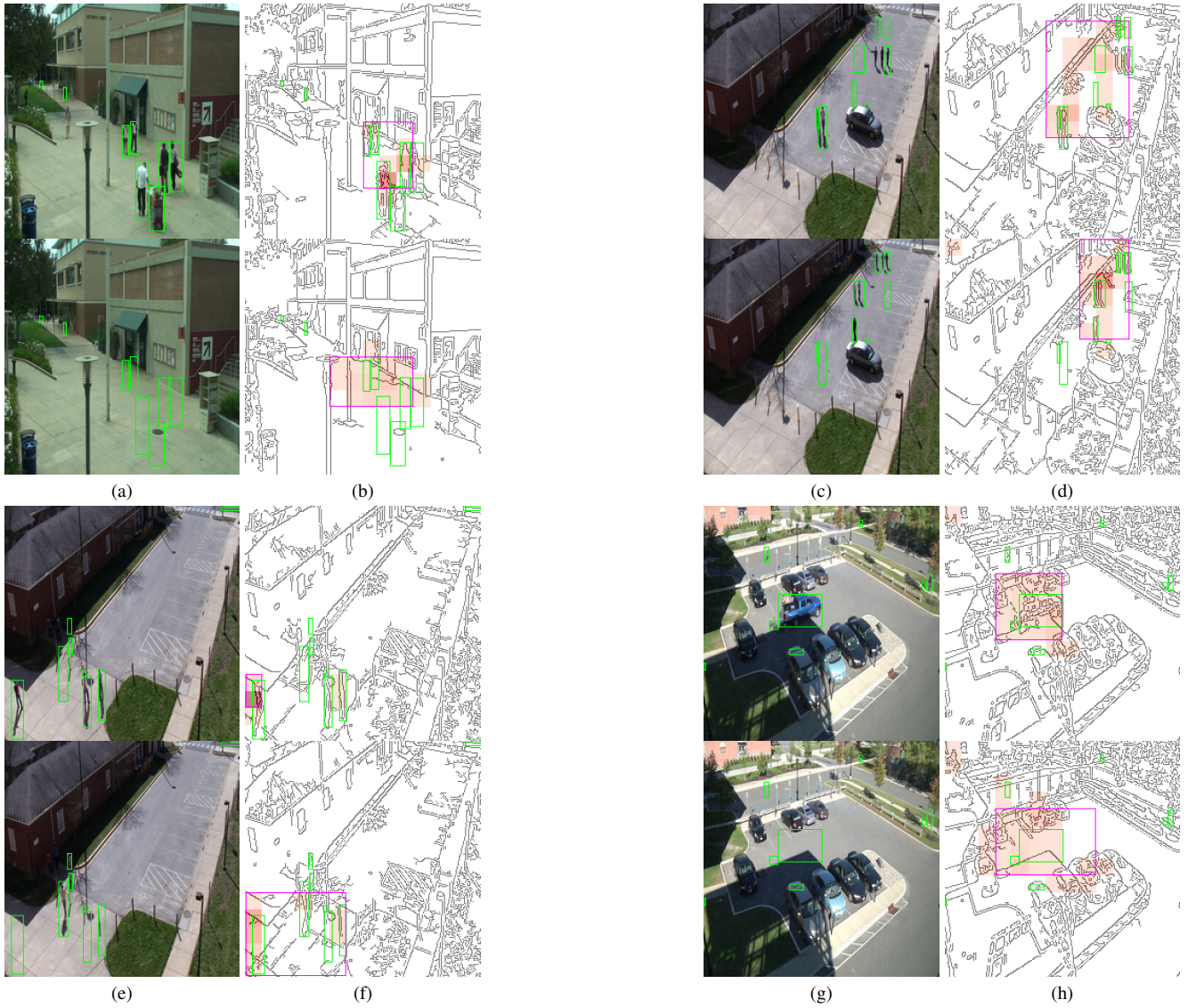


Figure A5. TAB's change localization on STD with B/16. STD contains images with multiple changes, which naturally leads to lower values in the attention maps.



## G. OpenImages-I additional results

### G.1. Correcting the attention map for change pairs

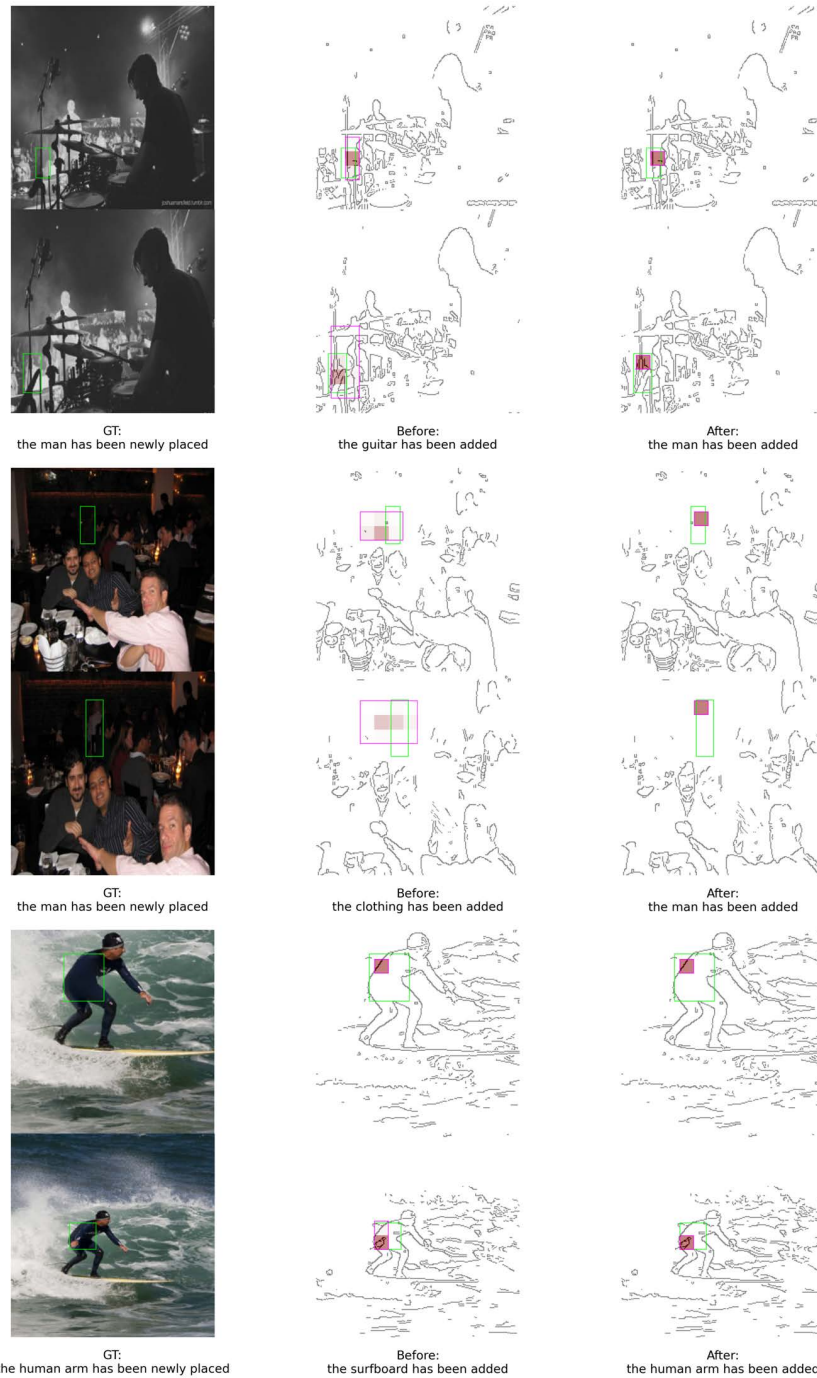
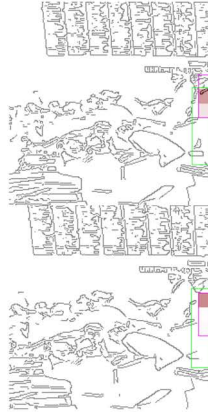


Figure A6. Editing the attention map in TAB with B/16 helps the VLM to caption the changes more accurately.



GT:  
the clothing is missing



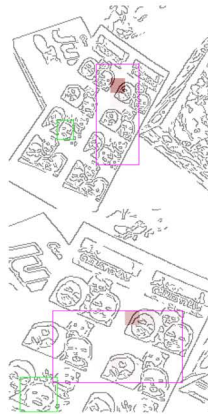
Before:  
the man is missing



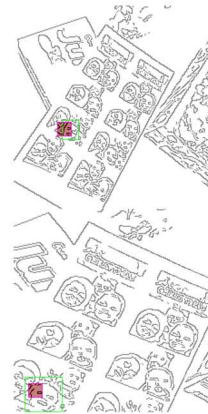
After:  
the clothing is missing



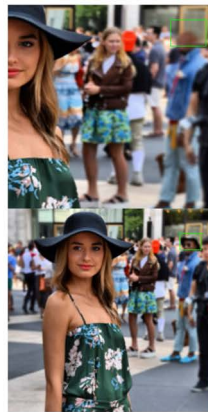
GT:  
the human head has appeared



Before:  
the mammal has been added



After:  
the human face has been added



GT:  
the fedora has been newly placed



Before:  
the human hair has been added



After:  
the sun hat has been added

Figure A7. Editing the attention map in TAB with B/16

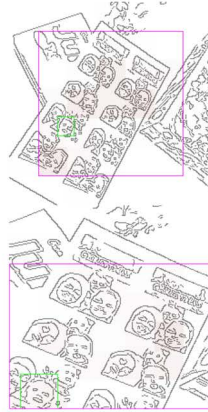


Figure A8. Editing the attention map in TAB with B/16





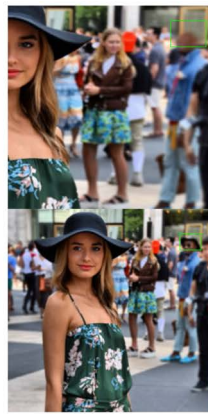
GT:  
the human head has been added



Before:  
the mammal has been added



After:  
the human face has been added



GT:  
the fedora has appeared



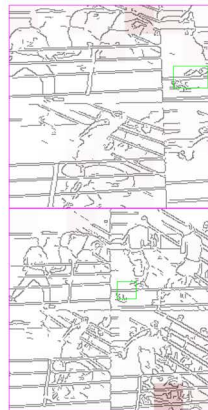
Before:  
the human hair has been added



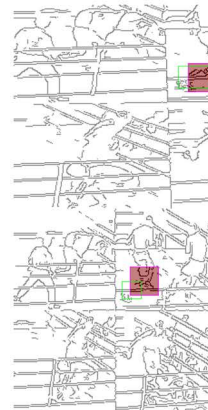
After:  
the sun hat has been added



GT:  
the sports equipment is gone



Before:  
the man has been added



After:  
the sports equipment is missing

Figure A9. Editing the attention map in TAB with B/32



Figure A10. Editing the attention map in TAB with B/32

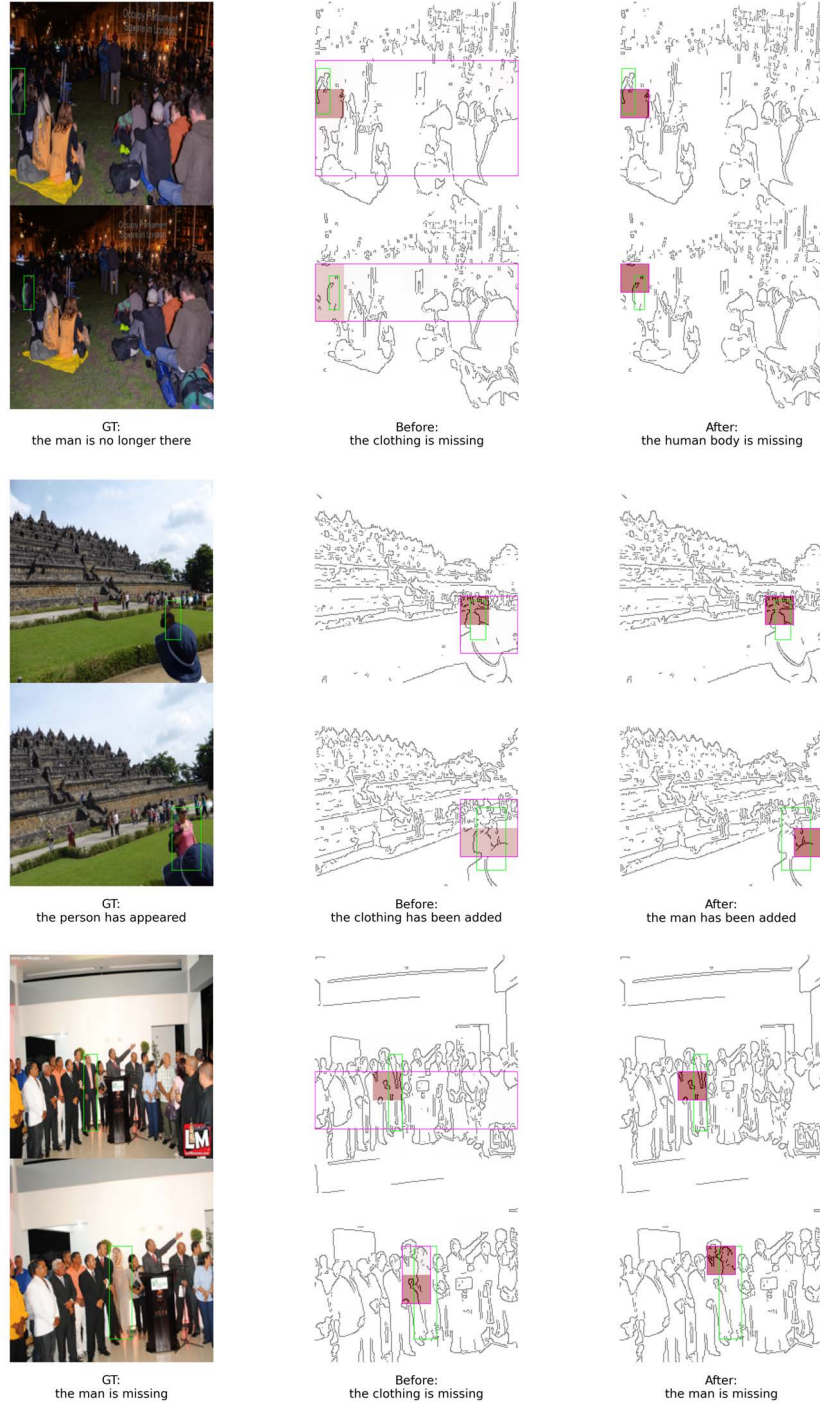


Figure A11. Editing the attention map in TAB with B/32.

## H. CLEVR-Change additional results

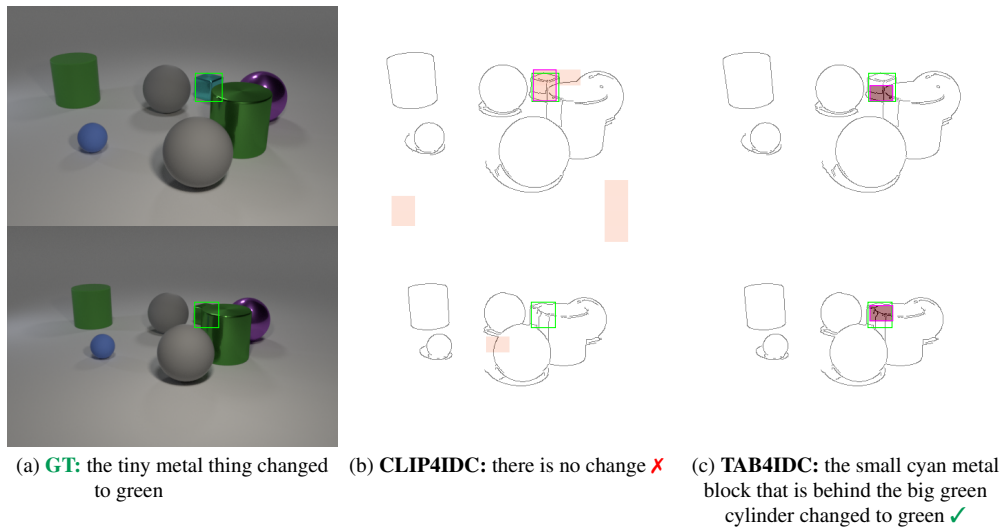

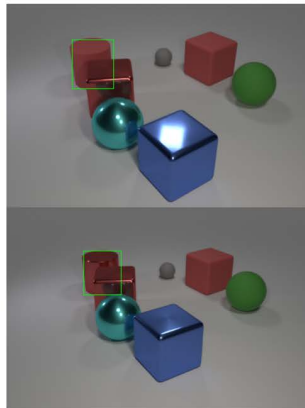
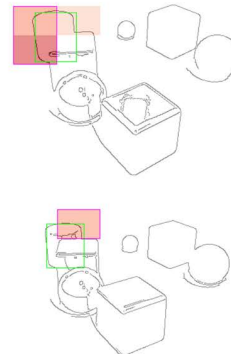


Figure A12. Compared to MHSA layer in CLIP4IDC (b) TAB better localizes the changed object that contributes to the predicted caption (c), for quantitative results we evaluate  $PG^+$  against the groundtruth ( $\square$ ) in .

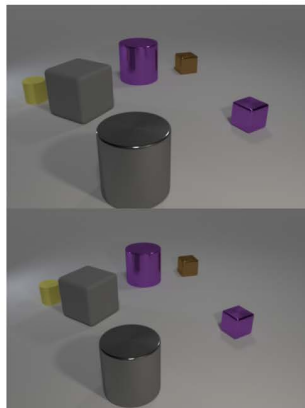
## H.1. Captioning and Localization



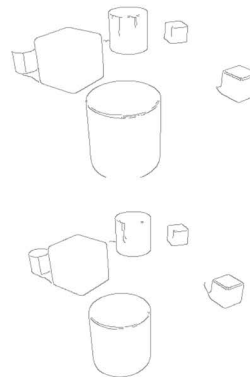
GT:  
the large matte cylinder turned shiny



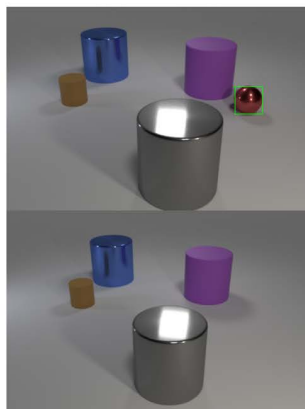
VLM:  
the large red rubber cylinder that is  
behind the big red rubber thing changed  
to metallic



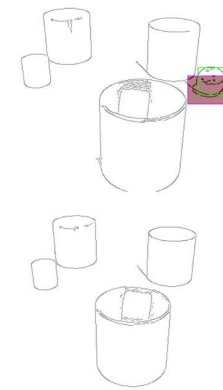
GT:  
the scene is the same as before



VLM:  
there is no change



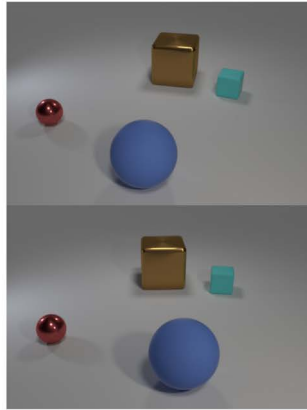
GT:  
the metal ball is gone



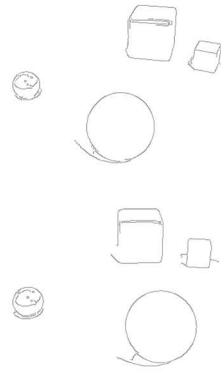
VLM:  
the small red metallic sphere that is  
right of the large gray cylinder is  
missing

Figure A13. Captioning: TAB4IDC with B/32

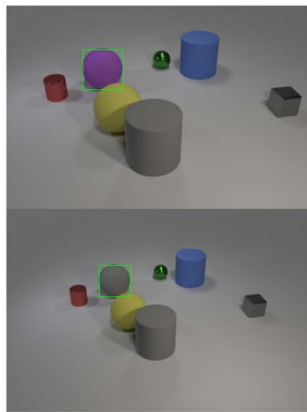




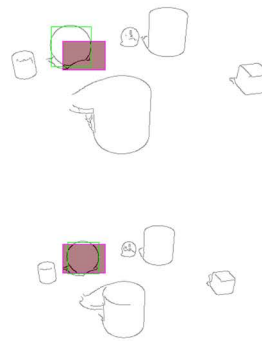
GT:  
the scene is the same as before



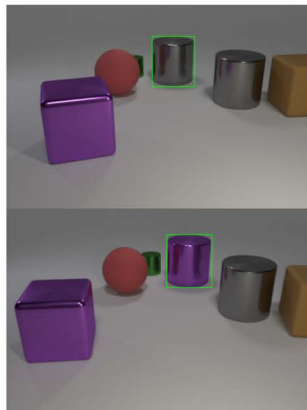
VLM:  
there is no change



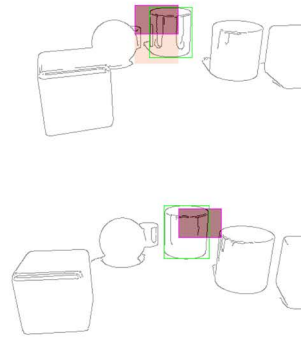
GT:  
the big purple object became gray



VLM:  
the purple rubber thing became gray

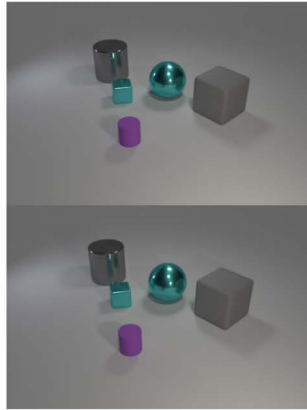


GT:  
the big gray metal cylinder behind the red matte object changed to purple

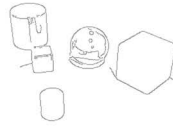
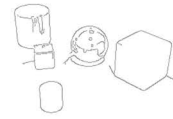


VLM:  
the big gray metallic cylinder that is behind the big brown rubber thing became purple

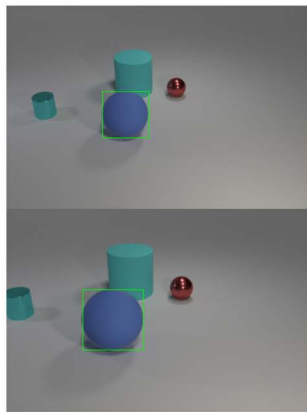
Figure A14. Captioning: TAB4IDC with B/32



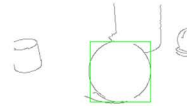
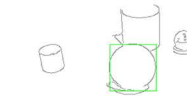
GT:  
the two scenes seem identical



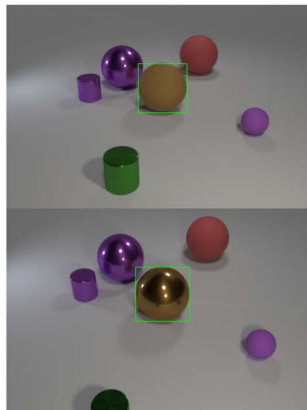
VLM:  
there is no change



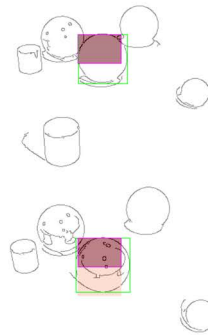
GT:  
the rubber sphere changed its location



VLM:  
there is no change

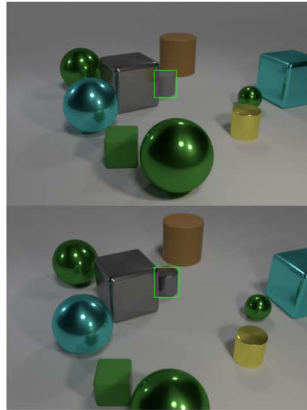


GT:  
the big brown ball changed to metallic

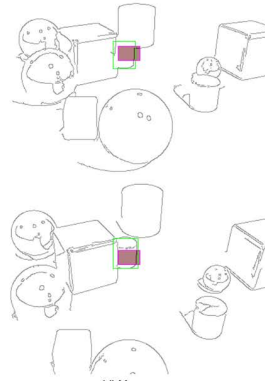


VLM:  
the large brown rubber sphere that is  
in front of the big purple rubber  
thing became shiny

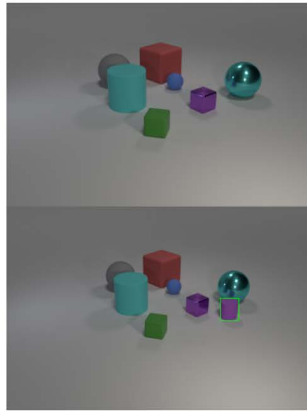
Figure A15. Captioning: TAB4IDC with B/32



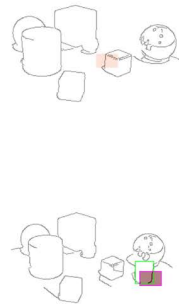
GT:  
the gray cylinder changed to shiny



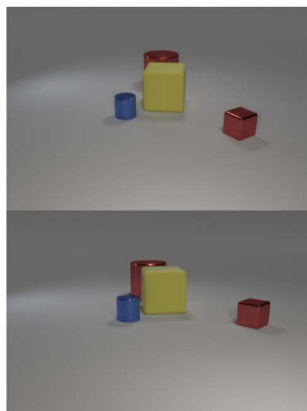
VLM:  
the tiny gray matte cylinder that is  
behind the small yellow cylinder changed to  
metallic



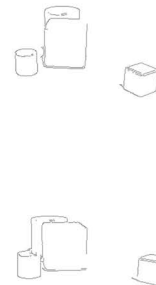
GT:  
the small cylinder has been newly placed



VLM:  
the tiny purple matte sphere that is  
to the right of the big green  
matte object has been added

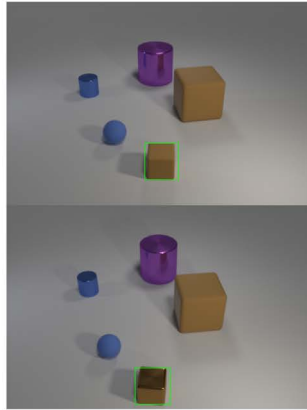


GT:  
the scene is the same as before

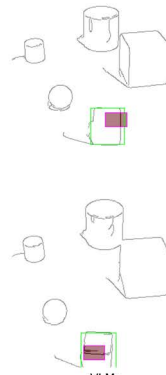


VLM:  
there is no change

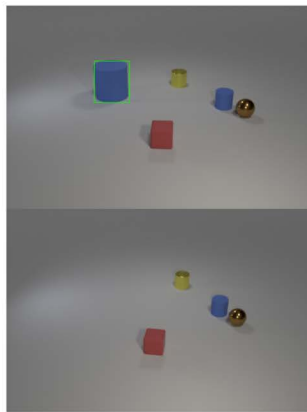
Figure A16. Captioning: TAB4IDC with B/16



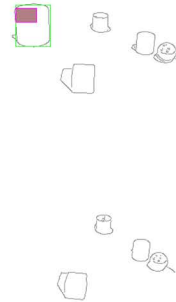
GT:  
the small brown cube became shiny



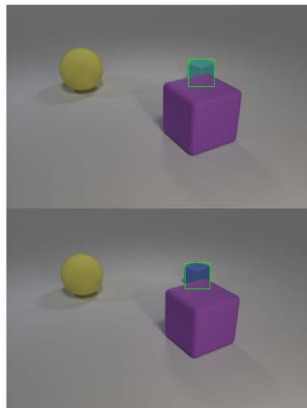
VLM:  
the tiny brown matte block that is  
in front of the small yellow matte  
object changed to metallic



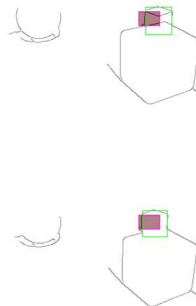
GT:  
the big matte cylinder is missing



VLM:  
the big blue matte cylinder that is  
left of the small blue matte object  
is gone

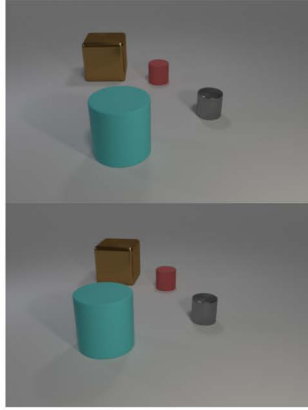


GT:  
the small object turned blue

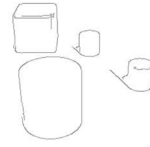


VLM:  
the tiny cyan matte block that is  
behind the small yellow cylinder changed to  
blue

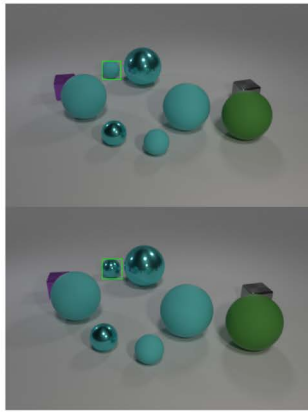
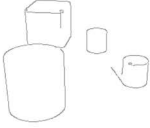
Figure A17. Captioning: TAB4IDC with B/16



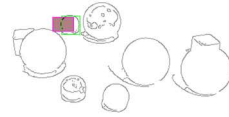
GT:  
nothing has changed



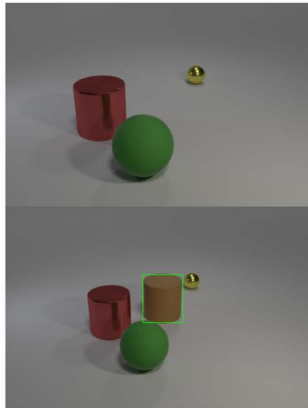
VLM:  
there is no change



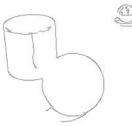
GT:  
the small cyan matte sphere behind the  
tiny gray metal cube became metal



VLM:  
the tiny cyan rubber ball that is  
behind the tiny yellow cylinder changed to  
metallic



GT:  
the rubber cylinder has been newly placed



VLM:  
the large brown matte cylinder that is  
behind the big red metal object has  
been added

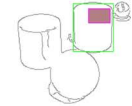


Figure A18. Captioning: TAB4IDC with B/16

## H.2. Correcting the attention map for change pairs

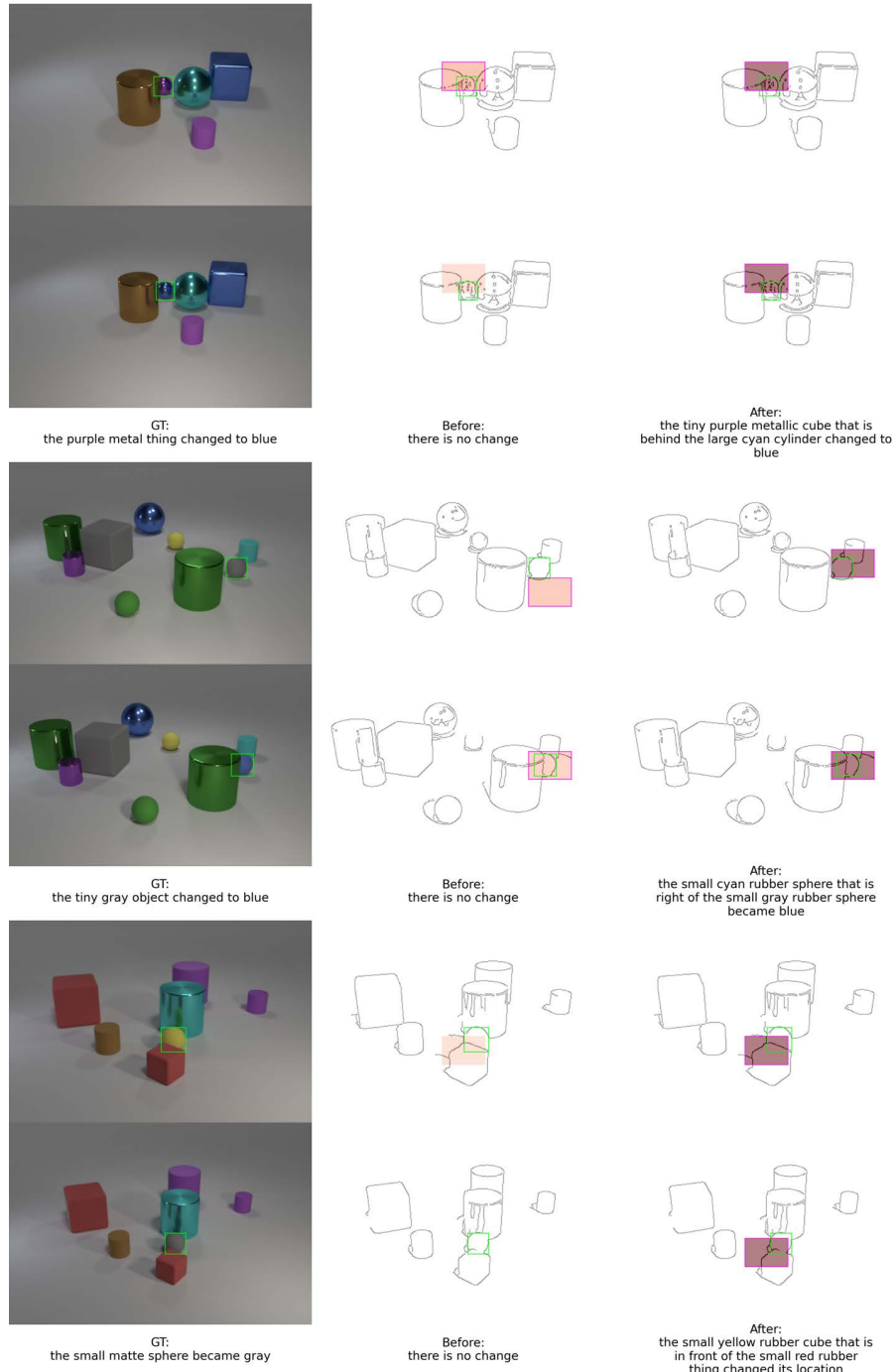


Figure A19. Editing the attention map in TAB with B/32

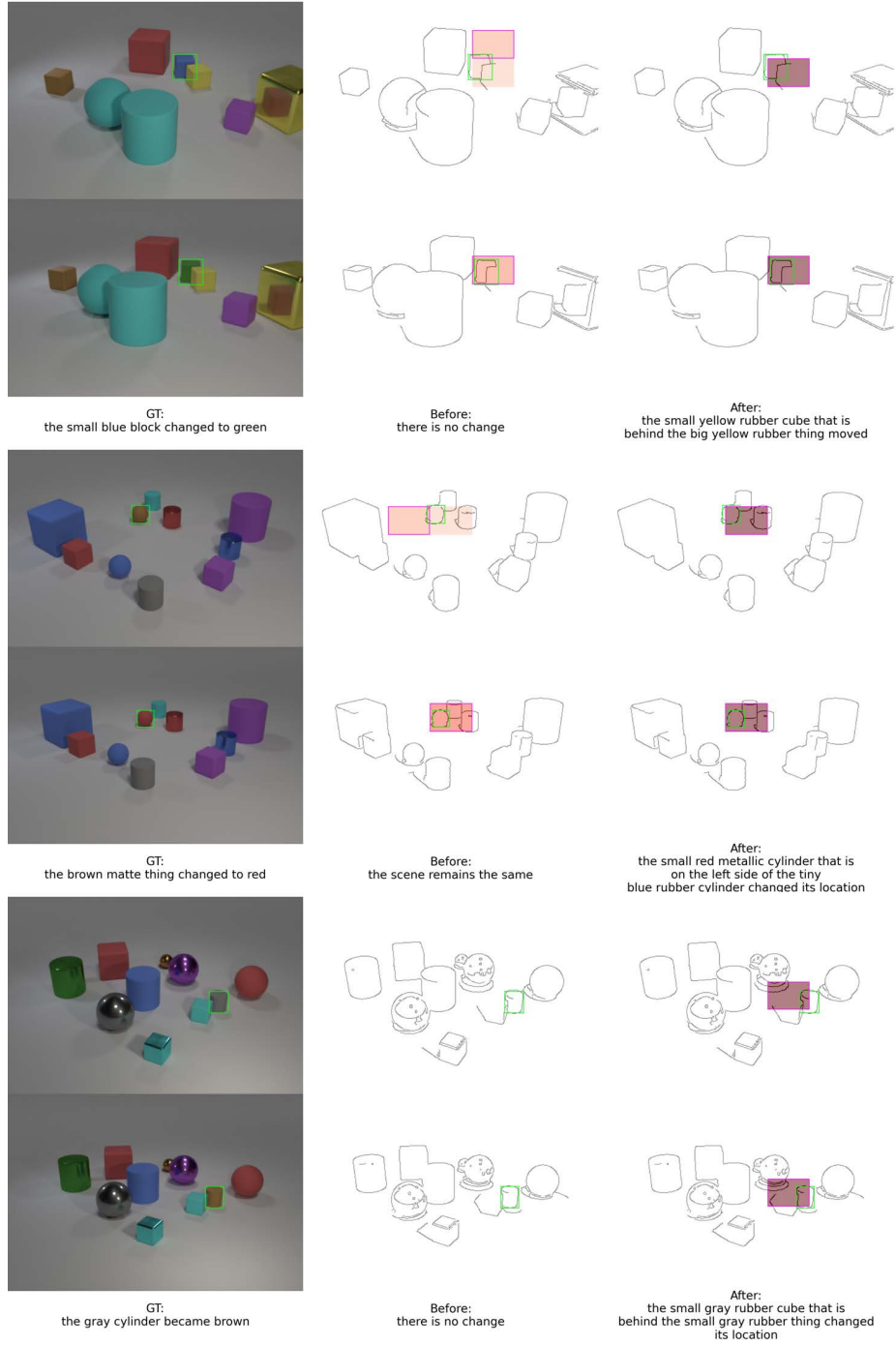


Figure A20. Editing the attention map in TAB with B/32

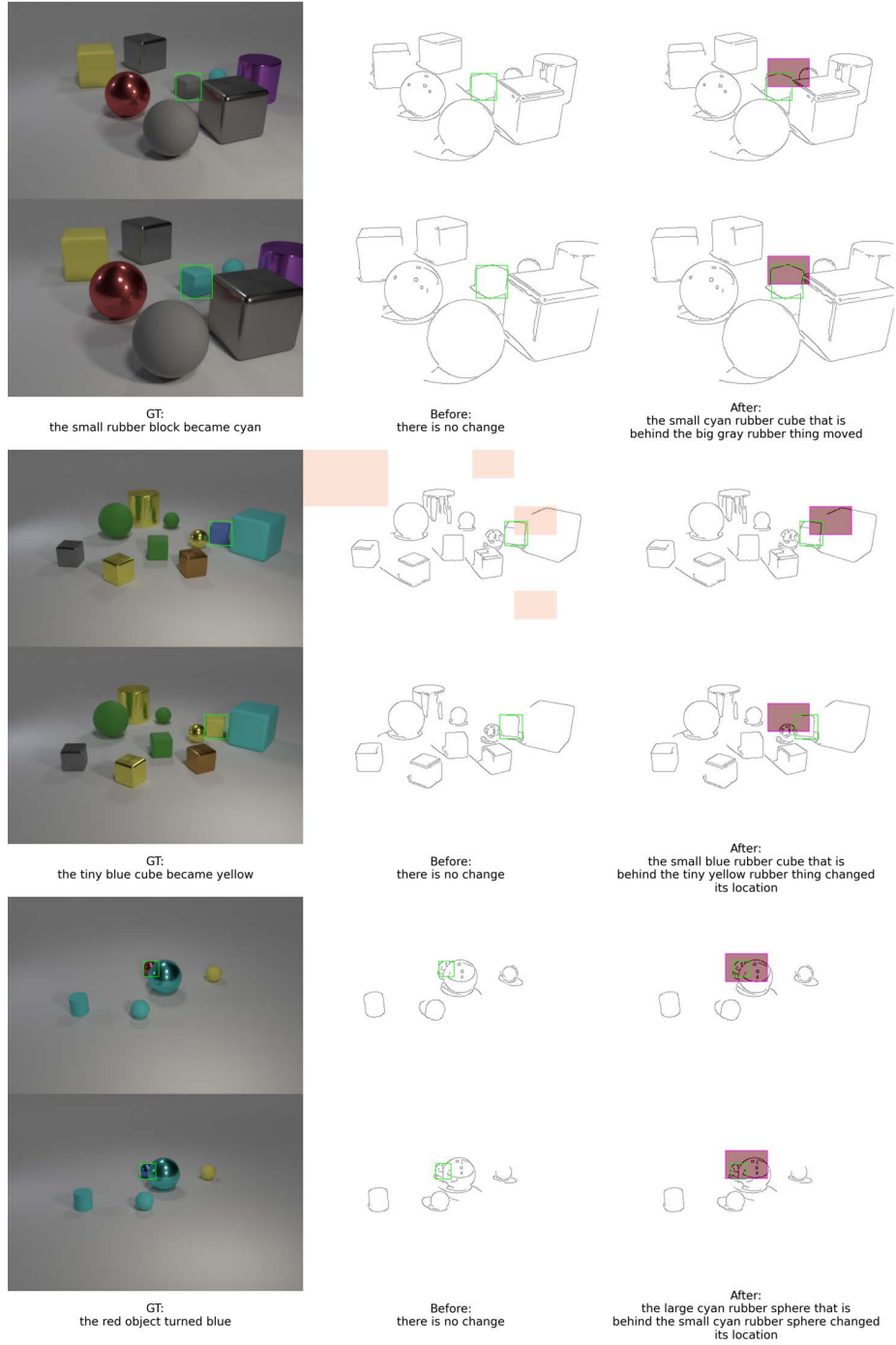


Figure A21. Editing the attention map in TAB with B/32



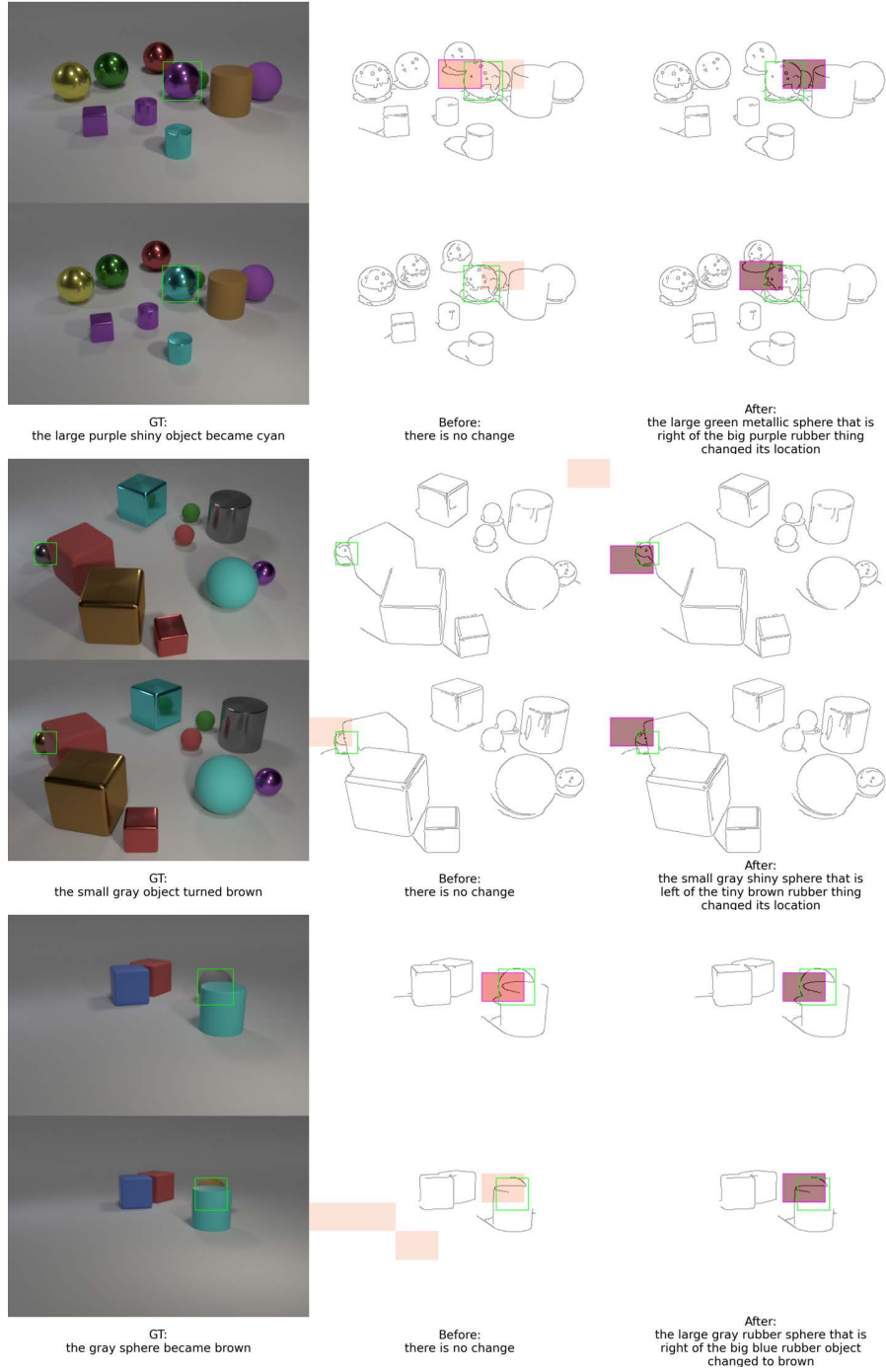


Figure A22. Editing the attention map in TAB with B/32

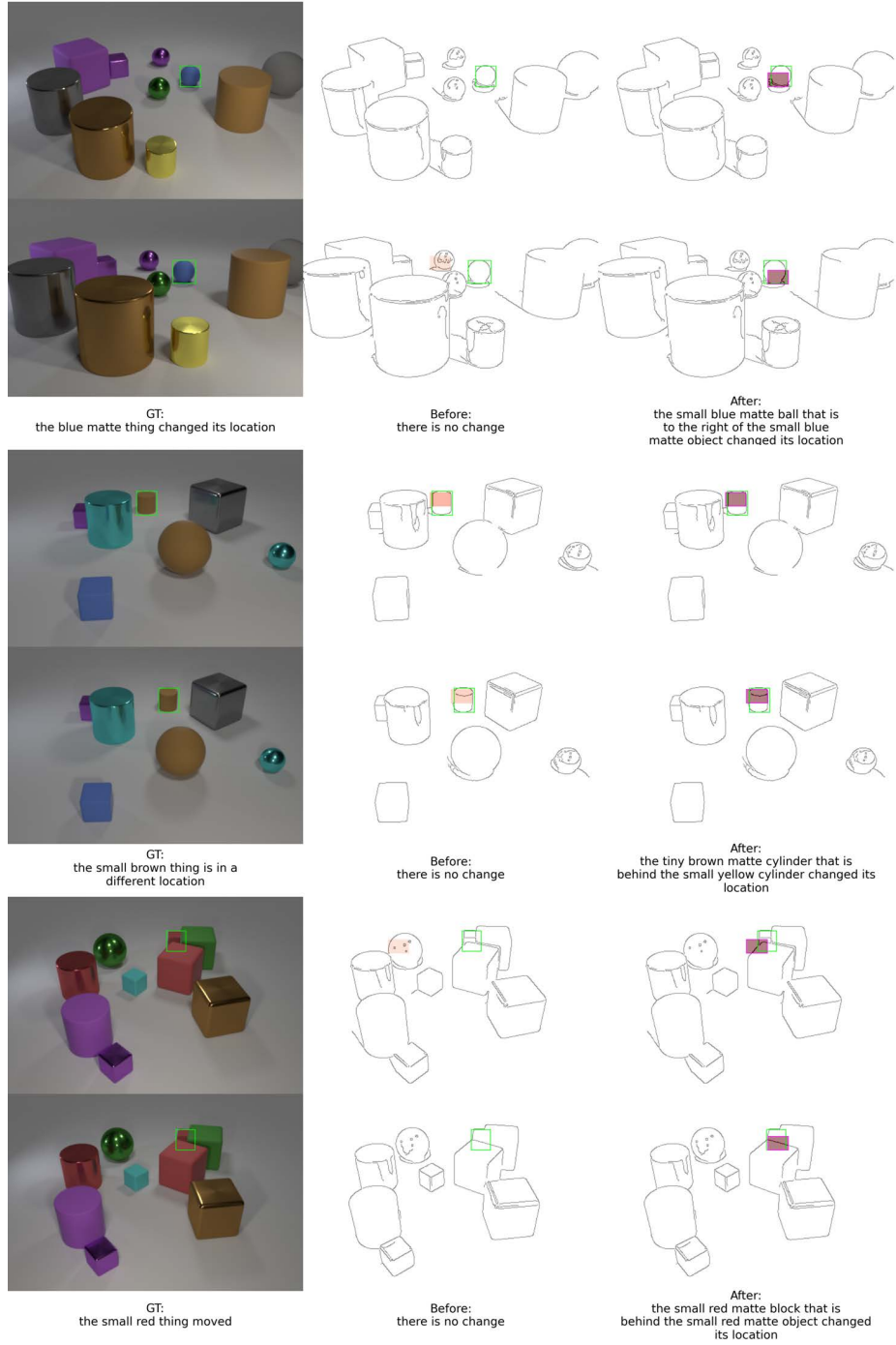


Figure A23. Editing the attention map in TAB with B/16

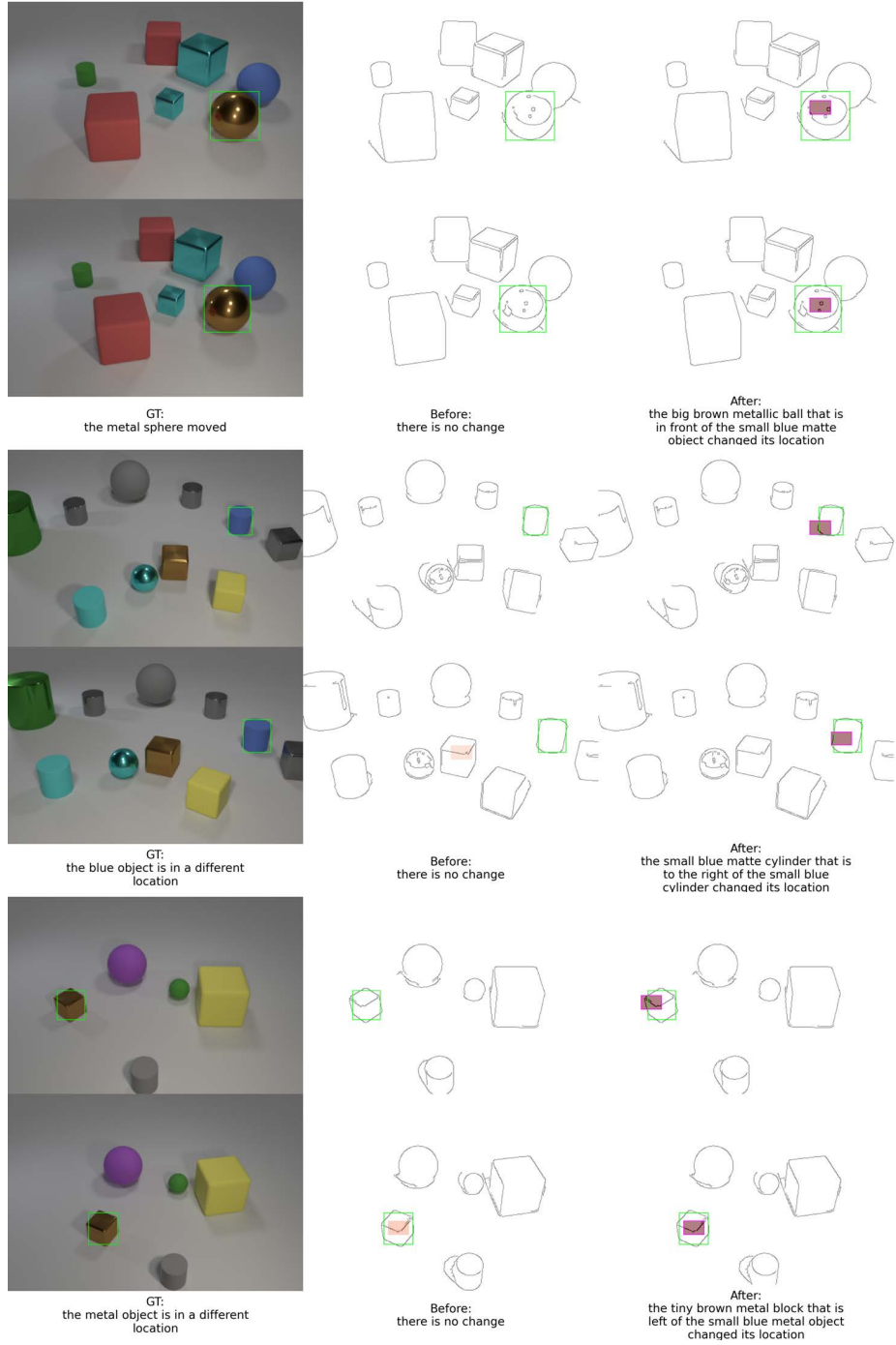


Figure A24. Editing the attention map in TAB with B/16

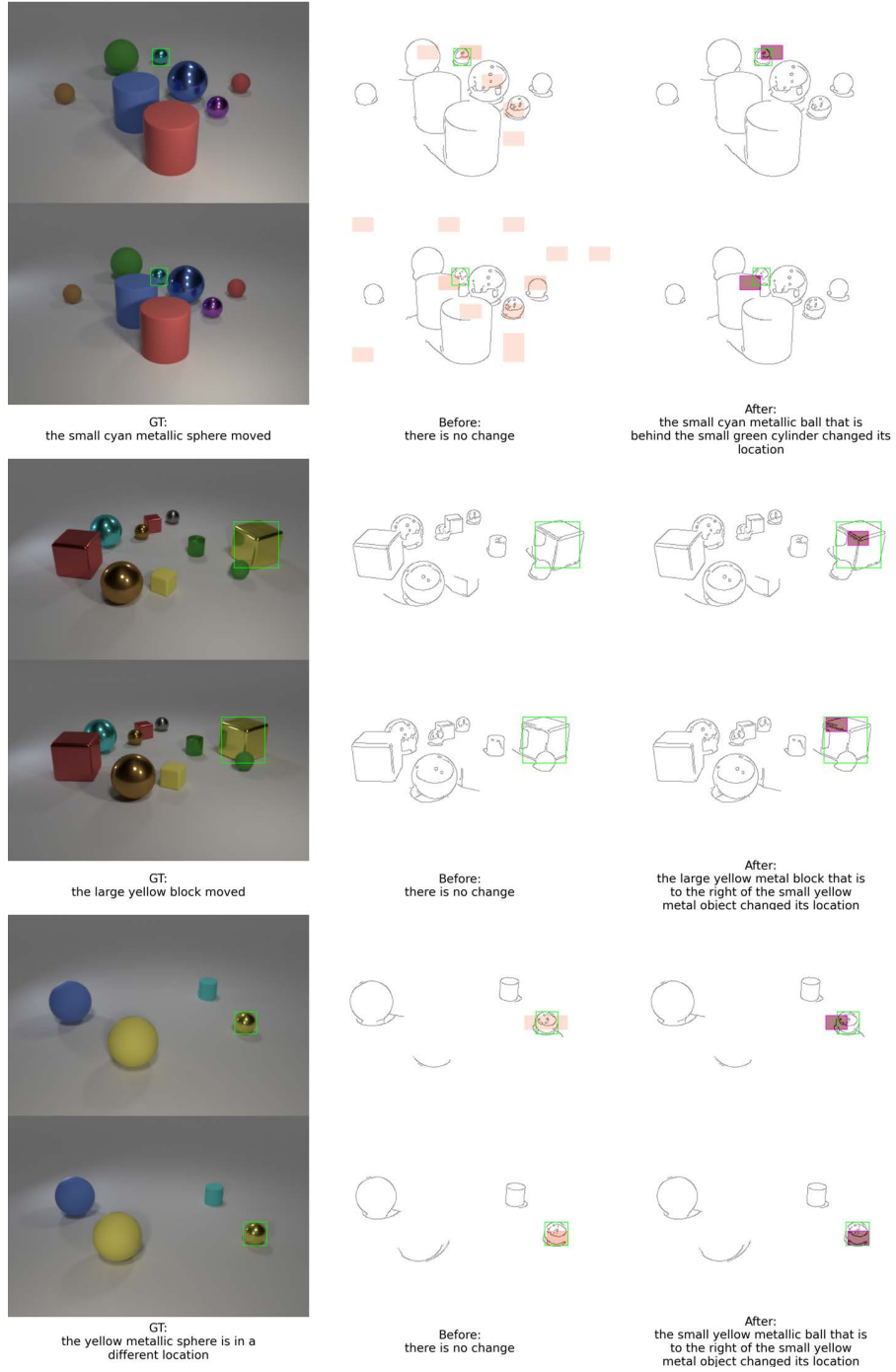


Figure A25. Editing the attention map in TAB with B/16

### H.3. Zeroing the attention map for no-change pairs

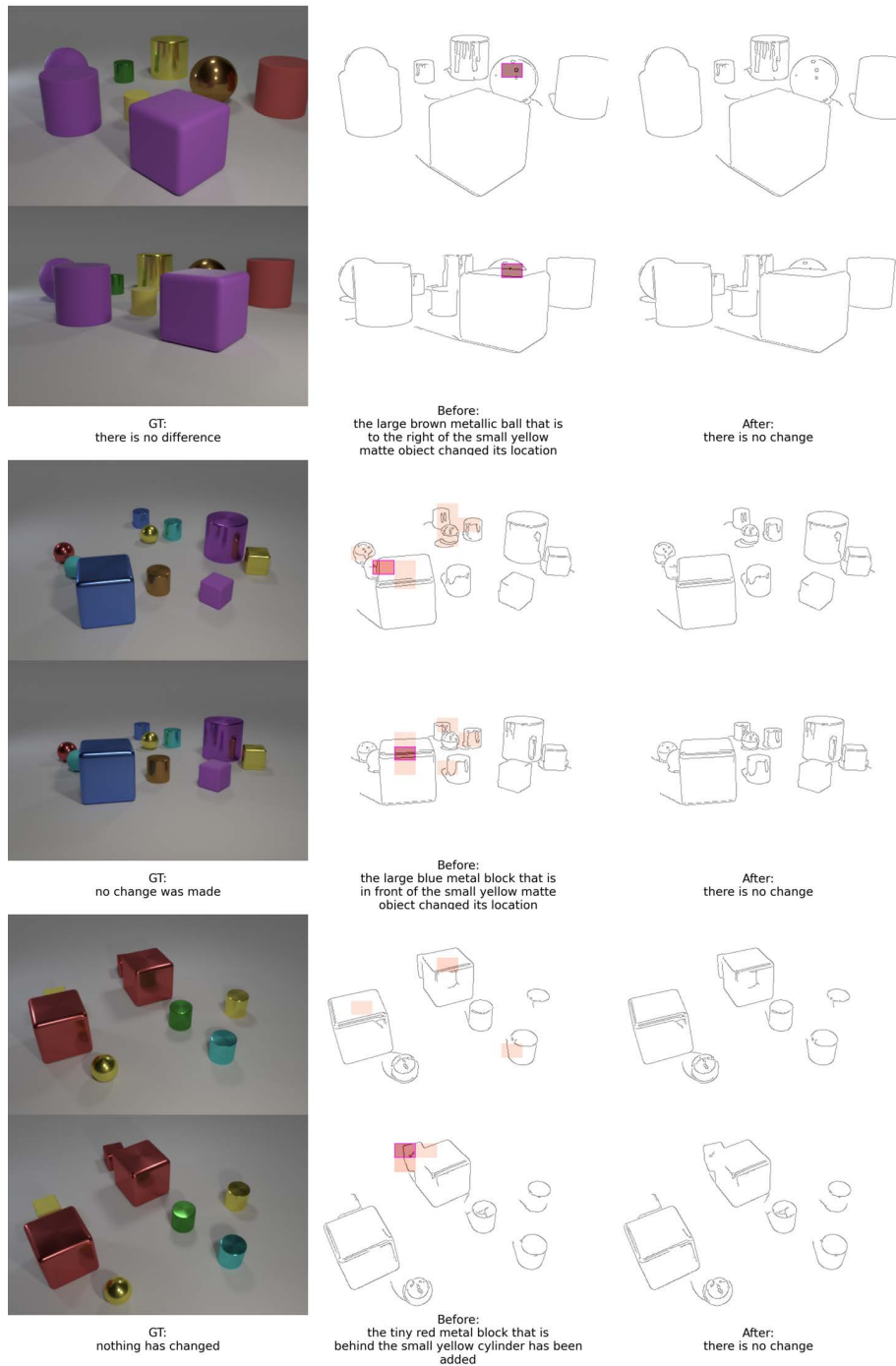


Figure A26. Editing the attention map in TAB with B/16

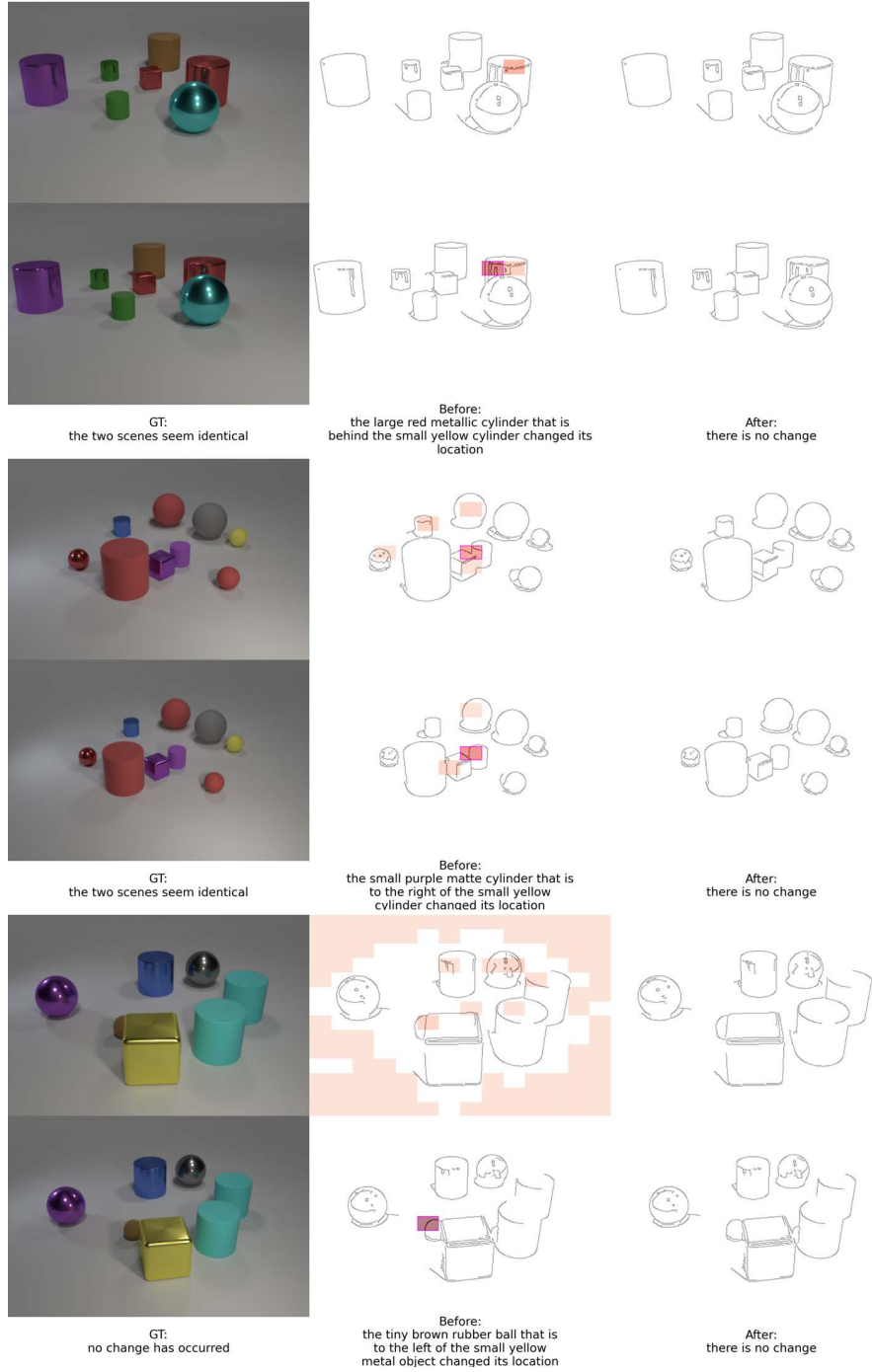
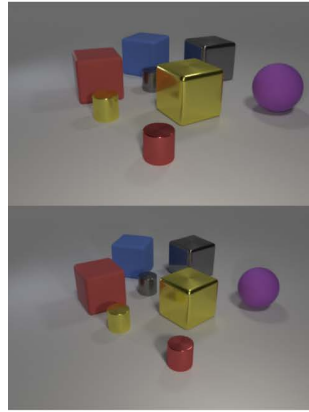
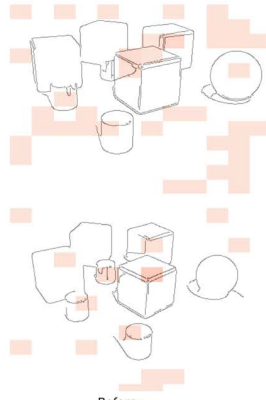


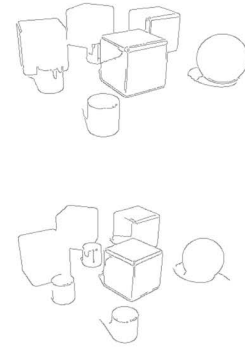
Figure A27. Editing the attention map in TAB with B/16



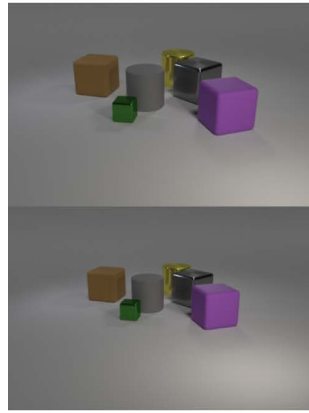
GT:  
nothing was modified



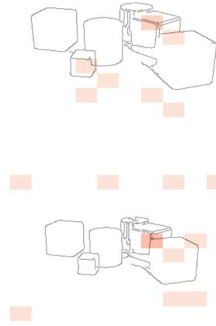
Before:  
the big gray metal block that is  
in front of the big yellow metal  
object changed its location



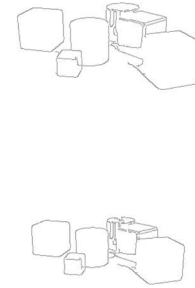
After:  
there is no change



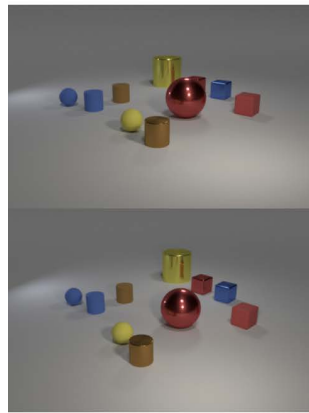
GT:  
there is no change



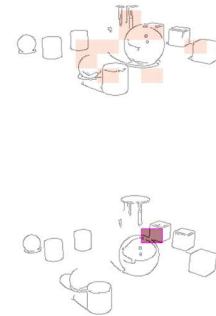
Before:  
the big gray metal block that is  
to the right of the big gray  
metal object changed its location



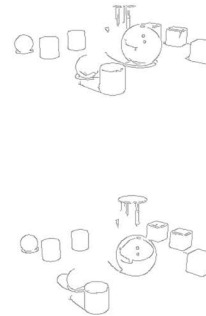
After:  
there is no change



GT:  
nothing has changed

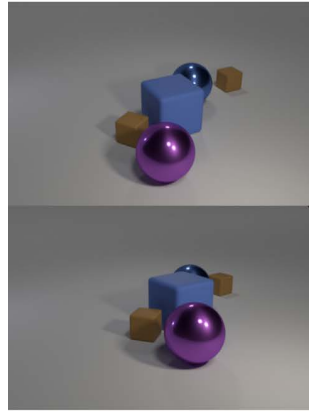


Before:  
the tiny red metal block that is  
behind the small blue matte object has  
been added

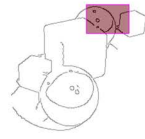


After:  
there is no change

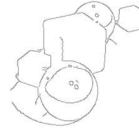
Figure A28. Editing the attention map in TAB with B/16



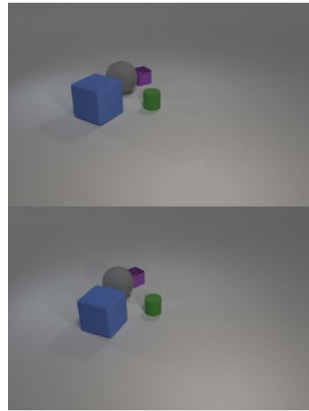
GT:  
there is no change



Before:  
the large blue metallic sphere that is  
behind the big blue rubber thing is  
missing



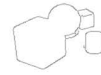
After:  
there is no change



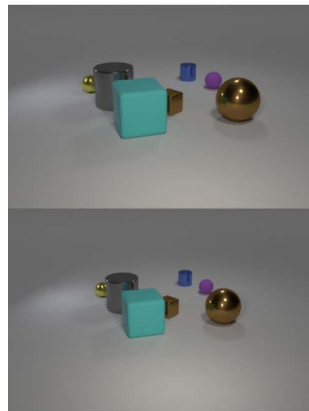
GT:  
there is no change



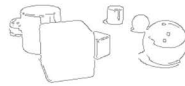
Before:  
the large gray rubber sphere that is  
behind the tiny green rubber thing is  
missing



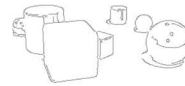
After:  
there is no change



GT:  
there is no difference



Before:  
the small brown metallic cube that is  
in front of the big yellow rubber  
thing moved



After:  
there is no change

Figure A29. Editing the attention map in TAB with B/16



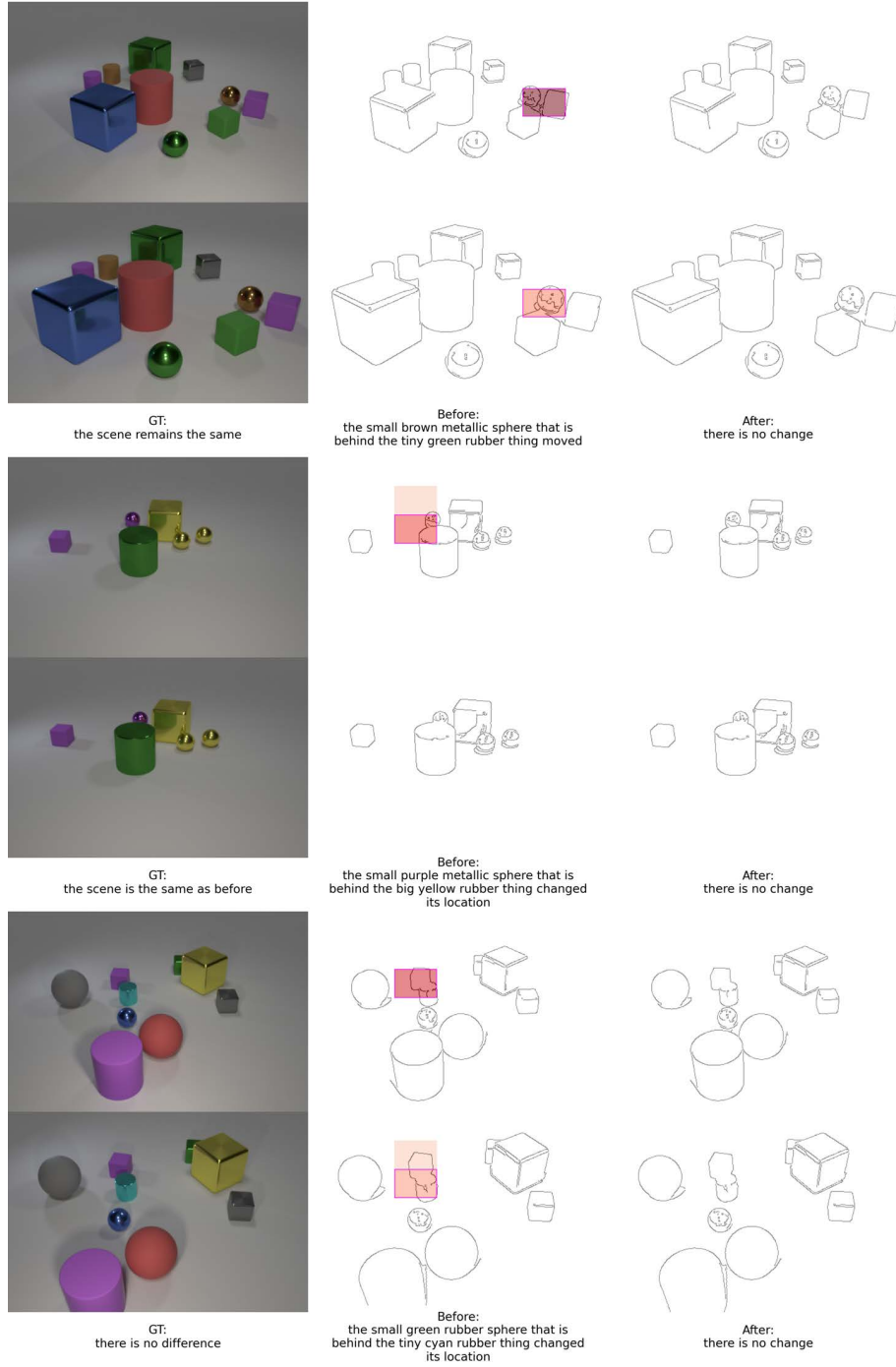
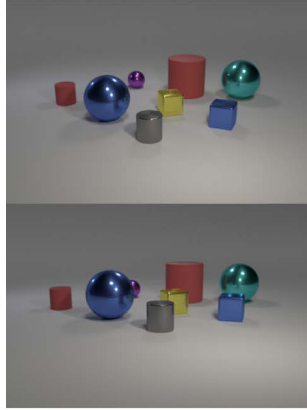
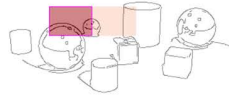


Figure A30. Editing the attention map in TAB with B/16



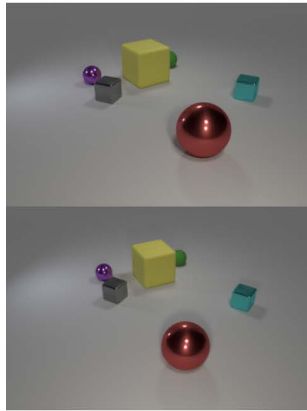
GT:  
the scene remains the same



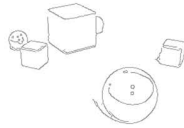
Before:  
the tiny purple shiny sphere that is  
behind the tiny blue rubber thing is  
missing



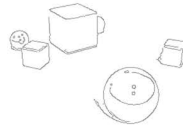
After:  
there is no change



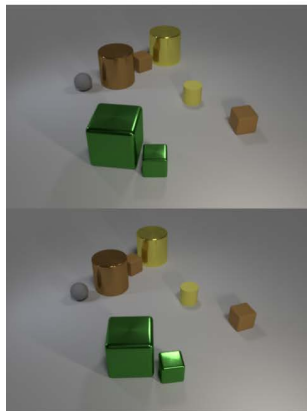
GT:  
there is no difference



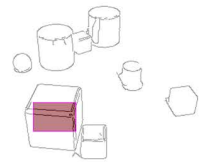
Before:  
the small green rubber sphere that is  
behind the big yellow rubber thing has  
been newly placed



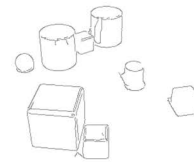
After:  
there is no change



GT:  
the scene is the same as before



Before:  
the large green metallic cube that is  
in front of the big yellow metallic  
thing is in a different location



After:  
there is no change

Figure A31. Editing the attention map in TAB with B/16